RCAAP

Repositório Científico de Acesso Aberto de Portugal

WP4 – D3 – RELATÓRIO SOBRE CKAN

JUNHO DE 20176

GOVERNO DE PORTUGAL | MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

FCT Fundação para a Ciência e a Tecnologia

Financiado por POS_Conhecimento | UNIÃO EUROPEIA FEDER

VERSÃO

Autor: José Carvalho; Filipe Furtado; Pedro Príncipe; Augusto Ribeiro

Contribuição: Eloy Rodrigues

Versão: 8.0

Distribuição: [Público]

Data de Criação: 25 de março de 2014

Última Atualização: 30 de novembro de 2016

GOVERNO DE PORTUGAL | MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA FCT Fundação para a Ciência e a Tecnologia

Financiado por POS_Conhecimento UNIÃO EUROPEIA FEDER

ÍNDICE

INTRODUÇÃO

Este documento visa analisar o sistema CKAN (*Comprehensive Knowledge Archive Network)* enquanto ferramenta de gestão de dados científicos no contexto do projeto RCAAP.
São analisados os aspetos técnicos, as funcionalidades, interoperabilidade e casos de uso comuns neste contexto nacional.

A análise da ferramenta teve como base a instância disponibilizada em: http://193.136.192.148/.

A FERRAMENTA CKAN

O sistema CKAN é uma ferramenta de gestão de dados que tem como principal objetivo tornar essa informação acessível para a comunidade através de ferramentas para publicação, partilha, pesquisa e reutilização da informação. Está focado para fornecedores de dados de vários âmbitos, quer seja, organizações, empresas, consórcios regionais ou instituições de nível nacional que pretendam disponibilizar e partilhar os seus dados.

Requisitos

Em termos de requisitos técnicos, depende do tipo de instalação que se pretende fazer, contudo, o recomendado é o uso do Ubuntu 16.04, 64-bit. Os pacotes necessários no servidor são: nginx apache2 libapache2-mod-wsgi libpq5.
O módulo WSGI deve estar ligado. Necessita também de PostgreSQL e SOLR.
No final da instalação vão existir duas aplicações a correr em simultâneo, o próprio CKAN e o DataPusher que é um serviço para importação automática de dados para o "DataStore extension". Para mais informações, consultar a página de apoio à instalação da aplicação.[1]

Comunidade

A comunidade do CKAN enquadra-se numa lista de distribuição para questões técnicas, exposição de problemas, bugs e respetivas correções. Existe também um grupo no Google Groups para discussão mais genérica da plataforma.[2]

Documentação

A documentação está disponível no website do projeto[3], e é composta por uma componente genérica de documentação da aplicação e outra focada na documentação para instalação. O projeto está também disponível no Github[4] e disponibiliza publicamente o seu roadmap.[5]

[1] http://docs.ckan.org/en/latest/maintaining/installing/install-from-package.html
[2] http://ckan.org/developers/mailing-lists/
[3] http://ckan.org/developers/docs-and-download/
[4] https://github.com/ckan/ckan
[5] https://waffle.io/ckan/ideas-and-roadmap

FUNCIONALIDADES DO CKAN

Publicar e Gerir Dados

Um interface intuitivo permite aos "dataset publishers" e aos curadores efetuar o registo na plataforma, atualizar e filtrar os "datasets" num modelo distribuído chamado 'Organizações'. As 'Organizações' permitem que cada um dos publicadores possa definir o seu conjunto de dados e procedimentos de aprovação, recorrendo a vários elementos. Esta situação implica que a responsabilidade de gestão e autorização dos registos pode ser realizada de forma distribuída, envolvendo departamentos, grupos de especialistas, etc. Neste caso a plataforma torna-se muito versátil e não existe obrigatoriedade de uma gestão centralizada.

Criar registos de dados no CKAN

Existem três procedimentos que podem ser adotados para criar os registos de dados no CKAN:

- Upload de ficheiros pelo interface web;
- Usando a API JSON do CKAN
- Através de procedimentos personalizados que permitem a importação de tabelas de dados (como o Google Spreadsheet).

Desta forma, a utilização mais comum será a utilização da interface web, contudo, os dados podem ser enviados através de JSON ou importação de folhas de cálculo.

Este processo permite a alimentação dos conjuntos de dados através de sistemas remotos, quer seja através de sistemas de medição que enviam a informação, quer através da exportação de CSV das aplicações e a sua posterior importação no repositório.

Agregação

Para efetuar regularmente a recolha de dados a partir de diferentes fontes, o CKAN desenvolveu um mecanismo que permite interpretar e importar os registos a partir de diferentes plataformas, tais como:

- Geospatial CSW Servers;
- Catálogos Web;
- Páginas indexadas na web ou diretórios acessíveis pela web;
- ArcGIS, Geoportal Servers e bases de dados acessíveis por Z39.50;
- Outras instâncias do CKAN.

Esta característica é a que define melhor a designação do sistema “Arquive Network”, permitindo um arquitetura distribuída de fontes de informação.

Workflow

Os dados podem ser públicos ou privados. Se forem privados estão exclusivamente disponíveis para os utilizadores autenticados numa organização. Os administradores podem aprovar os conjuntos de dados que podem ser tornados públicos ou não.

Pesquisa e descoberta de informação

O CKAN disponibiliza um interface de pesquisa muito semelhante ao que é considerado o “estilo google”, permitindo a pesquisa por palavra, delimitada ou não por aspas. Os utilizadores facilmente obtêm a lista dos conjuntos de dados disponíveis e podem realizar pesquisas nos conjuntos de dados. Para realizar a pesquisa nos conjuntos de dados temos as seguintes opções:

- Pesquisa dos campos em texto completo;
- Fuzzy-matching, é uma opção de pesquisa que permite encontrar os termos próximos daqueles que foram inseridos na pesquisa, em vez de encontrar termos exatos;
- Faceted search, permite uma pesquisa de navegação pelos campos disponíveis, permitindo ao utilizador filtrar os resultados passo a passo até encontrar o conjunto de dados pretendidos;
- Pesquisa pela API, todas as funcionalidade de pesquisa podem ser realizadas pela API do CKAN, facilitando a integração com outros sistemas.

Metadados

O CKAN disponibiliza um conjunto básico de metadados que permite descrever os conteúdos do documento, tais como:

- Título;
- Identificador único;
- Grupos;
- Descrição;
- Histórico de alterações do documento;
- Tipo de licença;
- Formato;

Contudo, apesar destes campos básicos, permite que sejam adicionados outros campos com elementos personalizados.

A API key do CKAN permite acesso e alteração dos metadados caso sejam dadas permissões para efetuar essas operações.

GeoSpatial

Sempre que a os dados sejam estruturados e inseridos com informação da localização, o CKAN pode apresentar os dados de uma referência geográfica num mapa interativo. Esta funcionalidade permite também pesquisa e descoberta de dados a partir de uma localização geográfica.

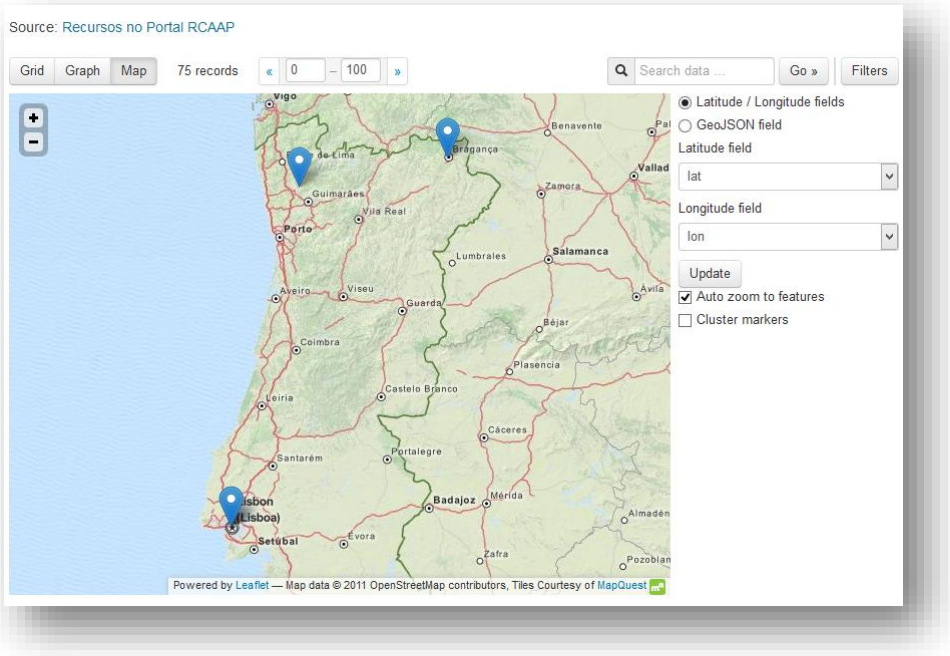


Figura 1 – Ilustração da funcionalidade GeoSpacial.

Interação com os utilizadores

O CKAN permite aos utilizadores adicionar comentários num conjunto de dados. Podem também promover ou partilhar a informação nas redes sociais, integrando com o Google+, o Twitter e o Facebook. Desta forma é fomentada a partilha e disseminação para a obtenção de indicadores de métricas alternativas.

Permite ainda "seguir" um conjunto de dados para receber informação sobre qualquer alteração ou nova atividade.

Visualização

O CKAN permite visualizar a informação em diferentes formatos de tabelas, nos formatos Excel ou CSV, facilitando o trabalho dos investigadores na análise dos dados. Estes mesmos dados podem ser também apresentados como um gráfico ou um mapa quando contêm informação de coordenadas geográficas.

Personalização

A interface Web do CKAN é facilmente customizada, existindo disponíveis procedimentos de customização no endereço[6].

Armazenamento

O armazenamento dos conjuntos de dados pode ser efetuado diretamente no CKAN ou numa localização externa, implicando esta segunda que seja disponibilizado a ligação para a localização onde foi feito o armazenamento. Os dados podem ser armazenados em qualquer formato e o CKAN disponibiliza ferramentas de visualização para alguns formatos mais comuns como CSV por exemplo.

Federação

Devido à funcionalidade de agregação disponível no CKAN, é facilitada a integração de dados, sendo uma mais-valia num cenário federado semelhante ao que temos atualmente para o projeto RCAAP.

Interoperabilidade

O CKAN disponibiliza uma RESTful JSON API para interagir com a base de dados. Esta API está presente em grande parte dos sistemas de repositórios e também em grande parte dos sistemas Integrados para Gestão de Bibliotecas, facilitando a interação dos sistemas existentes nas organizações que colaboram com o RCAAP. A documentação relativa à API do CKAN está disponível no endereço http://docs.ckan.org/.#sthash.KwoRIVPr.dpuf.

[6] http://docs.ckan.org/en/latest/theming.html

INTEROPERABILIDADE DO CKAN

O desenvolvimento de um serviço isolado pressupõe sempre a forma como o contexto onde é inserido opera. Neste contexto será obviamente considerado o contexto do projeto RCAAP e do sistema nacional de investigação científica.

Sendo desde sempre a interoperabilidade no projeto RCAAP um pressuposto basilar, segue-se uma análise às várias formas de interoperabilidade oferecidas pela ferramenta CKAN para o contexto nacional.

Apresenta-se de seguida um esquema das possibilidades de integração do serviço:

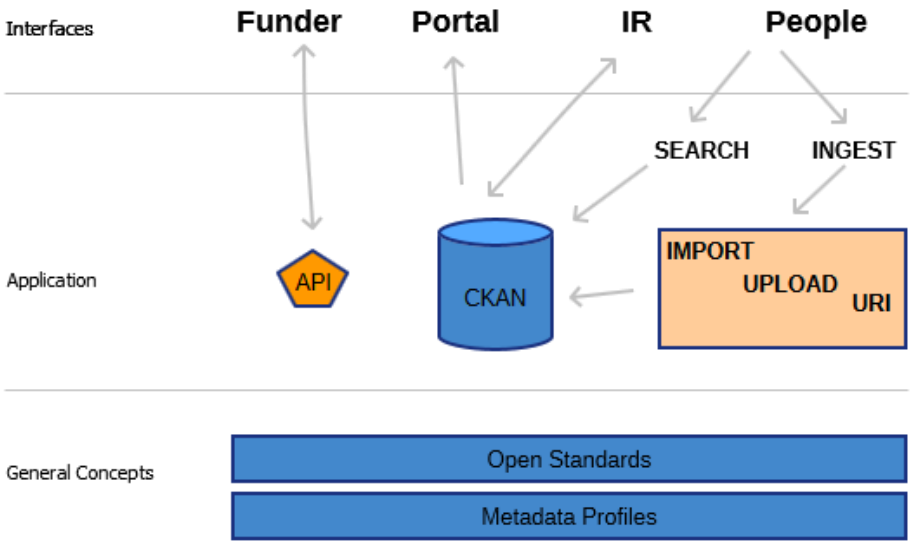


Figura 2 - Mapa de Interoperabilidade CKAN

Ao nível dos potenciais interfaces com o qual o CKAN poderá interagir devemos considerar por um lado as pessoas, quer do ponto de vista do investigador ou gestor dos dados, os repositórios institucionais para permitirem a referenciação de informação (do repositório de dados para o repositório de publicações, e vice-versa). Por outro lado temos o Portal RCAAP (ou outro tipo de agregadores) que permitem dar mais visibilidade a essa informação, partilhando e integrando essa informação noutros contextos de pesquisa. Finalmente, o contexto do financiador que tem em Portugal um papel relevante pois promove o depósito dos dados científicos financiados, quer seja pela FCT ou no programa H2020, e a sua associação às respetivas publicações científicas.

Ao nível da aplicação CKAN, podemos dividir em três partes. Por um lado temos o interface gráfico para pesquisa e submissão de informação orientado ao utilizador final, depois a aplicação CKAN e as suas funcionalidades, incluindo alguns módulos adicionais de funcionalidades e finalmente a API que permite criar um nível de interoperabilidade independente dos desenvolvimentos da aplicação CKAN.

Interface Humano

Considerando o interface humano da figura 2, o utilizador pode pesquisar a informação ou submeter novos conjuntos de dados de várias formas (envio ou referenciação dos conjuntos de dados).

Repositório Institucional

Ao nível de um repositório institucional de publicações científicas é possível:

- Agregar os conjuntos de dados no repositório
Desta forma toda a informação depositada no repositório de dados pode ser integrada também no repositório institucional para efeitos de pesquisa e relatórios.

- Referenciar os conjuntos de dados (dados <-> publicação)
Podem ainda ser referenciados os dados existentes no CKAN a partir do repositório institucional e vice-versa para que a relação entre os dados e as publicações seja mantida.

Portal RCAAP

Ao nível do agregador, é possível integrar o repositório de dados como um novo recurso, permitindo a pesquisa e recuperação dos registos existentes no repositório de dados.

Financiador

Ao nível do financiador, pode ser usada por exemplo a API para filtrar dos conjuntos de dados existentes no repositório, apenas aqueles que foram alvo de financiamento, tal como é desenvolvido já com as publicações.

Federação / Harvesting de Dados

Algumas instituições já possuem repositórios de dados próprios ou integrados nos repositórios institucionais com processos de gestão definidos. Nestes contextos será possível agregar informação remota no CKAN através das seguintes formas:

- Geospatial CSW Servers
- Catálogos Web
- Páginas HTML ou Web Accessible Folders
- ArcGIS, Geoportal Servers e Z39.50 databases
- Outras instâncias CKAN

Esta funcionalidade é utilizada no serviço data.gov[7] para incluir dados de várias agências. No caso do data.gov.uk[8] é utilizado para ir ao encontro da diretiva INSPIRE[9] (Infrastructure for Spatial Information in the European Community). É também utilizado pelo publicdata.eu[10] para incluir informação de diversos catálogos.

Esta última funcionalidade de agregação permite a criação de uma rede federada de repositórios de dados, permitindo criar um esquema semelhante ao Portal RCAAP com os repositórios institucionais, sendo neste caso possível depositar também no repositório central.

Este modelo permite por exemplo incluir repositórios de dados temáticos ou específicos a determinadas instituições com requisitos muito próprios. O CKAN usa o standard DCAT (Data Catalog Vocabulary), permitindo desta forma agregar outras fontes e formatos de informação.

API

Alinhado com as práticas atuais de disponibilização de interfaces para obtenção e disseminação de informação entre máquinas, o CKAN disponibiliza um API para acesso à informação dos conjuntos de dados.

[7] https://www.data.gov/
[8] https://data.gov.uk/
[9] http://inspire.ec.europa.eu/
[10] http://publicdata.eu/

A API permite:

- Consultar e pesquisar toda a informação constante no CKAN, quer dos metadados, quer do texto integral dos conjuntos de dados. Além disso, permite recuperar as ligações aos dados para que sejam referenciadas noutros sistemas.
- Listar e filtrar a informação existente.
- Acesso ao registo de alterações dos conjuntos de dados, também disponível via RSS e Atom.
- Aceder a estatísticas de uso como o número de downloads dos conjuntos de dados através da extensão do Google Analytics.
- Usar a versão RDF do catálogo, através da extensão RDF.
- Aceder a todo o conteúdo do CKAN no formato CSV e JSON

Esta API permite a leitura da informação mas também a atualização ou adição para utilizadores autorizados.

A documentação da API está disponível em http://docs.ckan.org/.

Extensões ao CKAN

Baseada no conceito de modularidade, o CKAN permite o desenvolvimento e/ou instalação de extensões ao código base, permitindo adaptar o sistema a contextos e necessidades distintas.
Existem já algumas extensões desenvolvidas e disponibilizadas pela comunidade e equipa de desenvolvimento: https://github.com/ckan/ckan/wiki/List-of-extensions.

Algumas extensões suportadas pela equipa que devem ser consideradas no processo de implementação:

- **ckanext-disqus** – Permite aos utilizadores comentar os conjuntos de dados através do serviço Disqus - https://disqus.com.
- **ckanext-googleanalytics** – Integra o serviço Google Analytics no CKAN, disponibilizando os downloads, rankings, etc.
- **ckanext-qa** – Verifica a ligação aos conteúdos, indica o índice de disponibilidade de informação e outras funcionalidade para evidenciar a qualidade do conjunto de dados.
- **ckanext-harvest** – Permite a importação de metadados de outras instâncias do CKAN.
- **ckanext-spatial** – Inclui funcionalidades relacionadas com dados georeferênciados, incluindo a possibilidade de pesquisar numa determinada localização.

- **ckanext-pages** – Permite adicionar páginas ao CKAN, à semelhança de um CMS básico.
- **ckanext-dashboard** – Apresenta a informação do CKAN num único dashboard.
- **ckanext-basiccharts** – Melhora a funcionalidade de criação de gráficos no CKAN.

Outras extensões não suportadas, desenvolvidas pela comunidade:

- **ckanext-oaipmh** – Agrega conjuntos de dados de um OAI-PMH e torna o CKAN um data provider com um interface OAI-PMH.
- **ckanext-shibboleth** – Autenticação Shibboleth para o CKAN - https://github.com/kata-csc/ckanext-shibboleth

Esquemas de Metadados

Ao nível dos metadados, o CKAN implementa nativamente, entre outros standards abertos,[11] o DCAT - Data Catalog Vocabulary[12] que permite descrever os conjuntos de dados através de RDF e desse modo facilitar a interoperabilidade com outros catálogos de informação. Cumpre ainda com o Data Catalog Interoperability Protocol.[13]

[11] http://ckan.org/open-standards/

[12] http://www.w3.org/TR/vocab-dcat/

[13] http://spec.datacatalogs.org

Comunidade Piloto

A funcionalidade do CKAN foi testada com um conjunto de dados fornecidos a título de exemplo pelo Campo Arqueológico de Mértola (ver figura 3).

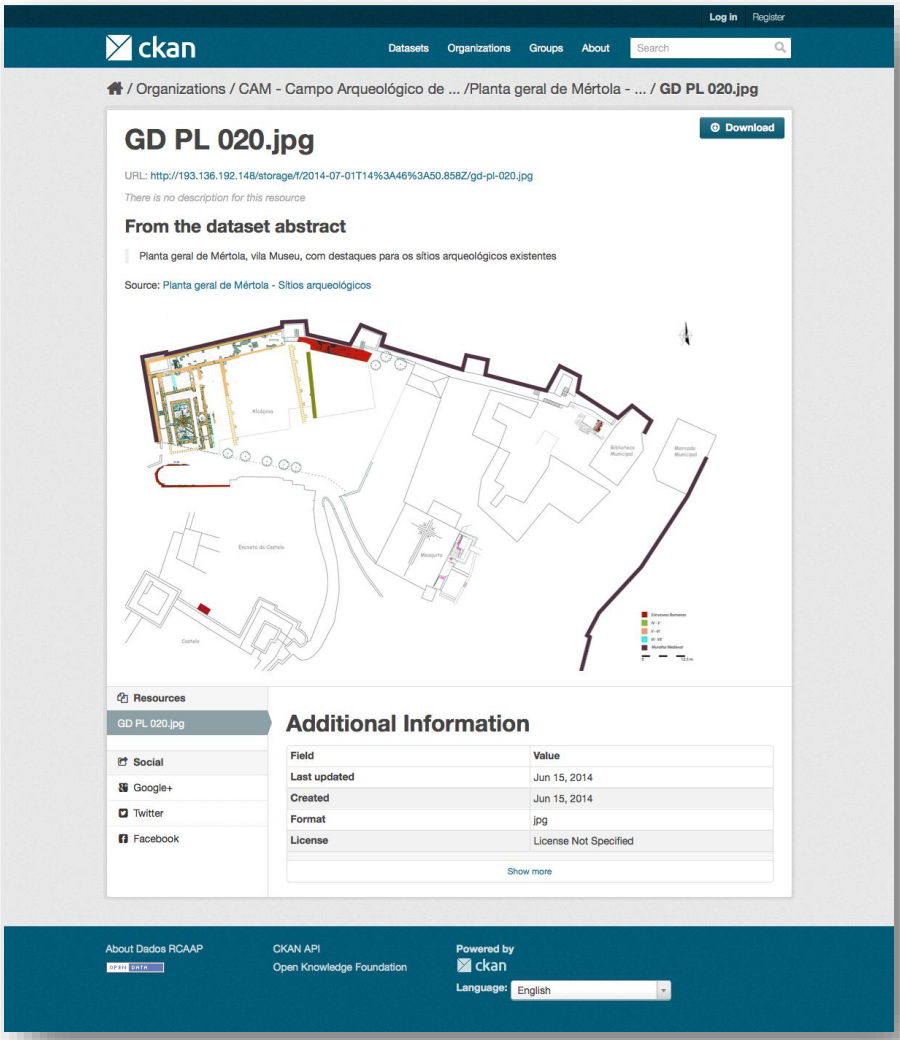


Figura 3 – Frontend do sistema CKAN, ilustrando os dados fornecidos pelo Campo Arqueológico de Mértola.

A tipologia de trabalhos testados foram mapas, fotografias e tabelas de dados. Foi identificado inicialmente um modelo de dados para a descrição dos itens mas não foi implementado na aplicação. O processo de submissão foi manual através do carregamento e descrição pelos colaboradores do CAM e em pouco tempo conseguiram disponibilizar vários exemplos de conjuntos de dados com uma descrição básica.

Conclusões sobre análise da aplicação CKAN

O CKAN é um sistema open-source, altamente customizável, escrito em Python e Javascript (web-frontend), particularmente usado em portais de dados governamentais, tendo também já sido implementado em repositórios de dados de investigação.[14,15] Tendo em vista estas duas aplicações (portal e/ou repositório de dados de investigação) apresentam-se em seguida algumas vantagens e desvantagens deste sistema. Esta lista baseia-se no artigo online[16] e na publicação de Amorim *et al*.[17]

Vantagens:

- algoritmo de procura fuzzy matching
- suporta versionamento de documentos, uma característica muito útil no que toca à curadoria de digital de dados de investigação
- possui uma poderosa API, altamente costumizável
- representação RDF de metadados, o que facilita a integração em sistemas Open Linked Data, como por exemplo o portal VIVO
- funcionalidade comprovada no use-case de dados governamentais, por exemplo: datahub.io, catalog.data.gov, data.gov.uk, data.gov.au, entre outros
- previsão gráfica dos dados depositados em formato de folha-de-cálculo, compatível com geo-localização
- fornece um ambiente de colaboração com comentários e rating para os respetivos conjuntos de dados
- desenhado originalmente como agregador de dados publicados noutros repositórios[18] torna-o particularmente indicado como fonte de dados de investigação do portal RCAAP

Desvantagens:

- não foi desenvolvido especificamente para preservação digital de dados
- não suporta nativamente outros esquemas de metadados, pressupondo a criação de novos plugins/*add-ons*, para cada novo esquema
- não suporta de base a criação de um identificador permanente (DOI), mas é possível fazê-lo.

[14] http://www.ckan.org/instances/
[15] http://dataportals.org/
[16] http://www.edawax.de/2013/09/adapting-ckan-for-open-research-data
[17] http://link.springer.com/article/10.1007/s10209-016-0475-y
[18] http://eprints.lincoln.ac.uk/9778/1/CKANEvaluation.pdf

- a funcionalidade de previsão gráfica (folhas de cálculo) será útil apenas para uma fração dos dados primários de investigação.
- atualmente, é utilizado em poucos repositórios de dados (não se encontra listado na lista de software de repositórios de dados http://opendoar.org/find.php?format=charts) o que claramente indica e implica uma comunidade limitada.
- escrito em Python, o que pode originar algumas dificuldades técnicas na integração na rede RCAAP, principalmente baseada em DSpace (Java).
- sistemas como DSpace, ou ePrints claramente abordam "use-cases" relativos a curadoria de dados, focando-se também na preservação, o que é reconhecido como uma característica central para uma infraestrutura de um repositório de dados de investigação.

Adicionalmente, o projeto Orbital[19] identificou as seguintes vantagens do sistema CKAN:

- integração com o ambiente de repositório institucional (interface com sistemas CRIS, repositórios institucionais, DMPOnline, armazenamento em rede)
- integra o contexto e atividade do processo de investigação
- boa gestão diferenciada de acessos, com diferentes níveis de privilégios administrativos, por exemplo para parceiros de investigação externos
- boas e abrangentes ferramentas de pesquisa
- adesão a normas Open Linked Data (com formato RDF)
- backup e armazenamento em acesso partilhado (por exemplo, via Dropbox)
- linha de comandos e boa interface web, para depósito e atualização de dados
- URIs permanentes para a citação, por exemplo, DOIs
- Importação e exportação de formatos de dados comuns
- interligação de conjuntos de dados (por projeto, tipo, output científico, pessoa, etc.)
- gestão e licenças de dados
- suporte comercial; um sistema amplamente usado pela comunidade

Da mesma forma, foram identificadas as seguintes desvantagens ou questões nas quais o projeto focou o desenvolvimento do sistema:
- modelo segurança: necessidade de desenvolvimento de diferentes níveis de acesso, autenticação.
- implementação do conceito de "Projetos", sendo este amplamente usado pela comunidade académica.

[19] http://orbital.blogs.lincoln.ac.uk/2012/09/06/choosing-ckan-for-research-data-management/

- adaptação da terminologia CKAN à terminologia académica e desenvolver documentação nesse sentido
- o upload, edição ou alteração de ficheiros em lote, está somente disponível através da linha de comandos
- falta de integração com outros sistemas ownCloud/Dropbox, e protocolos (SWORD2)

TAREFAS NECESSÁRIAS PARA ENTRADA EM PRODUÇÃO

No sentido de integrar o CKAN como um serviço para a comunidade, sugerem-se algumas alterações no sistema:

1. **Tradução do Interface** – Não existe um interface traduzido para português, e poderia ser definido como contribuição da comunidade para este serviço: http://docs.ckan.org/en/latest/contributing/i18n.html

2. **Alteração do Layout Gráfico** – Tal como acontece com outros serviços disponibilizados pelo projeto RCAAP, a imagem do projeto deve ser mantida uniforme para que a associação do serviço seja automática ao projeto: http://docs.ckan.org/en/latest/theming/index.html

3. **Configuração de Cronjobs** – Será ainda necessário configurar tarefas agendadas para o envio de notificações por email: http://docs.ckan.org/en/latest/maintaining/email-notifications.html

4. Configuração do **Google Analytics** para obtenção de indicadores de acesso ao repositório

5. Definição de **métricas para monitorização** do serviço (Nº de Datasets; Nº de Instituições;…).

6. Definição, descrição e procedimentos do **tipo de serviço** prestado.

7. Desenvolvimento do *add-on* compatível (ou configurações) com o **esquema DataCite**[20].

8. Inclusão de **identificadores persistentes** (DOI)

[20] https://www.datacite.org/

CASOS DE USO

Após a análise aplicacional, podemos abordar a materialização deste serviço na comunidade nacional que poderá ter perspetivas e análises diversas, conforme o contexto. Designamos aqui algumas possibilidades com base na figura 4 onde são visíveis três diferentes níveis.

O primeiro, o nível institucional ("*Institutional Context*") identificamos os fornecedores de dados que podem ter diversos sistemas e serviços para partilha de dados (repositórios em diversas plataformas e serviços ou APIs), podendo ter um carácter institucional ou temático.

Num segundo nível temos um portal de dados científicos (*"Data Portal"*) que serve para agregar todos os recursos do nível anterior, permitindo a existência de um único ponto de pesquisa e monitorização de dados científicos.

Finalmente, o Portal RCAAP ("Research Portal") que agrega os dados e os integra para efeitos de pesquisa com as publicações científicas.

Os serviços que podem ser disponibilizados pelo projeto RCAAP respeitam este contexto e são descritos de seguida.

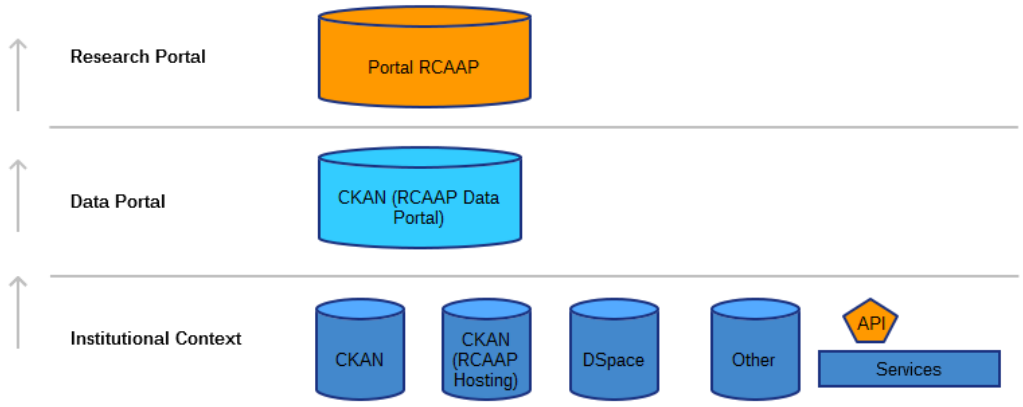


Figura 4 - Esquema de Tipos de Serviços

C1 – Contexto de Serviço Centralizado e Partilhado no RCAAP (SaaS)

No contexto do projeto RCAAP, poder-se-ia desenvolver um conceito de repositório de dados científicos, multidisciplinar que poderia integrar diretamente os conteúdos, com as devidas restrições com base no nível de serviço, assim como conteúdos existentes nas instituições que seriam apenas referenciados no Repositório de Dados (agregados) [21].

Recomenda-se ainda que o serviço em si esteja integrado e exista um acompanhamento constante por parte da equipa. Este contexto seria semelhante ao atual serviço de Repositório Comum, mas com a possibilidade adicional de integrar conteúdos externos. Este modelo foi já testado no primeiro projeto piloto, disponível em http://dados.rcaap.pt/.

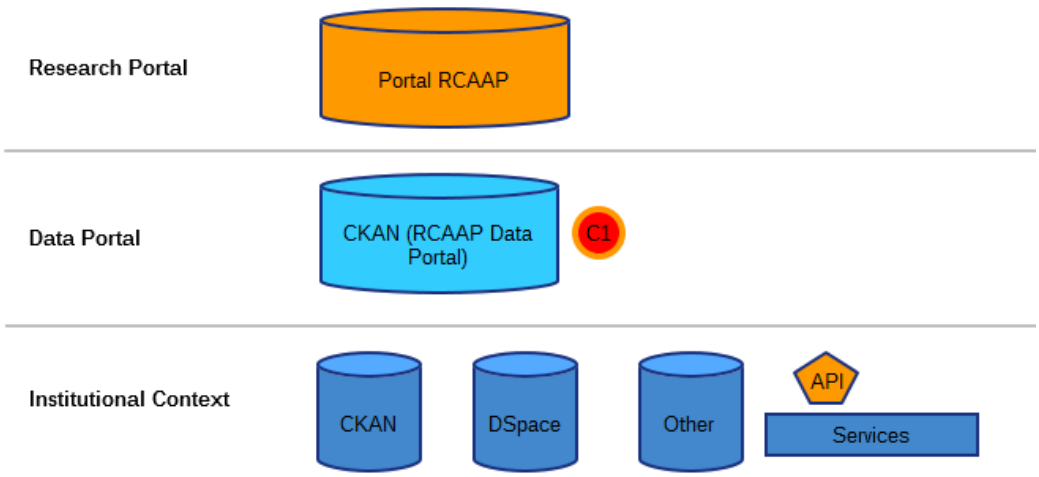


Figura 5 - Contexto 1 - Serviço Centralizado e Partilhado no RCAAP

[21] http://ckan.org/features-1/federate/

C2 – Contexto Centralizado com Serviço de Alojamento (SaaS)

Neste contexto, há ainda a possibilidade de conjugar um repositório central de dados (Data Portal) e permitir o alojamento **em regime de SaaS para repositórios de dados locais,** tal como já para os repositórios DSpace no SARI ou revistas científicas com o SARC.

A grande vantagem do repositório central é a possibilidade de incluir como fornecedor de dados serviços que alimentam o portal através de APIs por exemplo assim como a possibilidade de integrar centralmente comunidades com pequenas dimensões (semelhante ao conceito adotado no Repositório Comum).

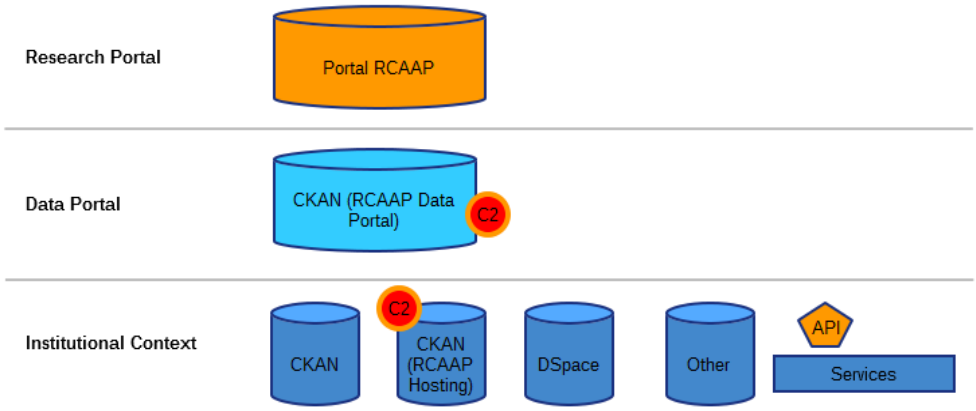


Figura 6 - Contexto 2 - Serviço de Alojamento de Repositórios de Dados Científicos com "Data Portal"

C3 – Contexto de Utilização Institucional (SaaS)

Neste terceiro cenário considera-se apenas o serviço de alojamento de repositórios de dados sem a intermediação do “Data Portal”. Neste caso, os dados dos repositórios locais são integrados diretamente no Portal RCAAP ou “Research Portal”.

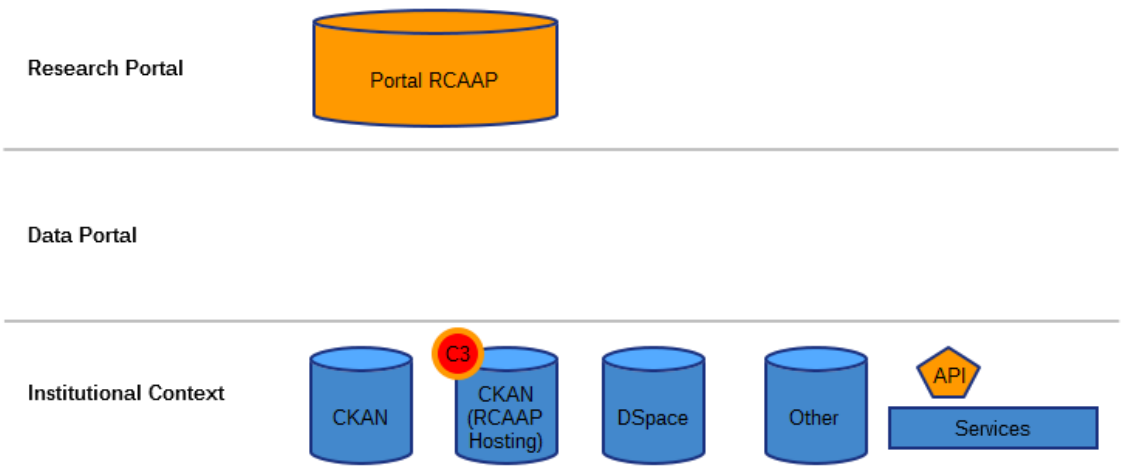


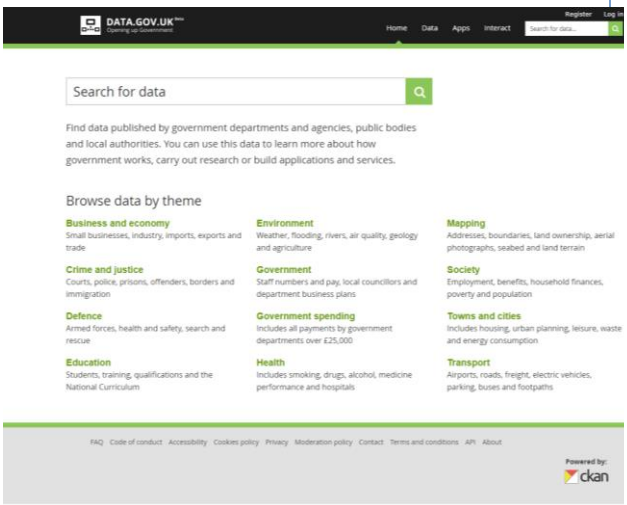
Figura 7 - Contexto 3 - Serviço de alojamento de Repositório de Dados Científicos

BOAS PRÁTICAS DE UTILIZAÇÃO

Neste capítulo identificam-se e descrevem-se casos de uso do sistema CKAN de acordo com os três contextos identificados para o projeto RCAAP: o contexto C1 (Serviço Centralizado Partilhado) e C2/3 (Serviço de Alojamento e de Utilização Institucional) definidos, no capítulo anterior.
Estes correspondem, respetivamente, ao uso do sistema enquanto portal de dados abertos governamentais (C1) e enquanto repositório de dados de investigação (C2/3).

O sistema CKAN deverá ser usado como plataforma modular e generalista, fazendo uso da sua potencialidade enquanto sistema interoperável, como descrito na figura 2. Esta característica deverá ser utilizada para garantir a agregação de dados de investigação de diferentes repositórios.

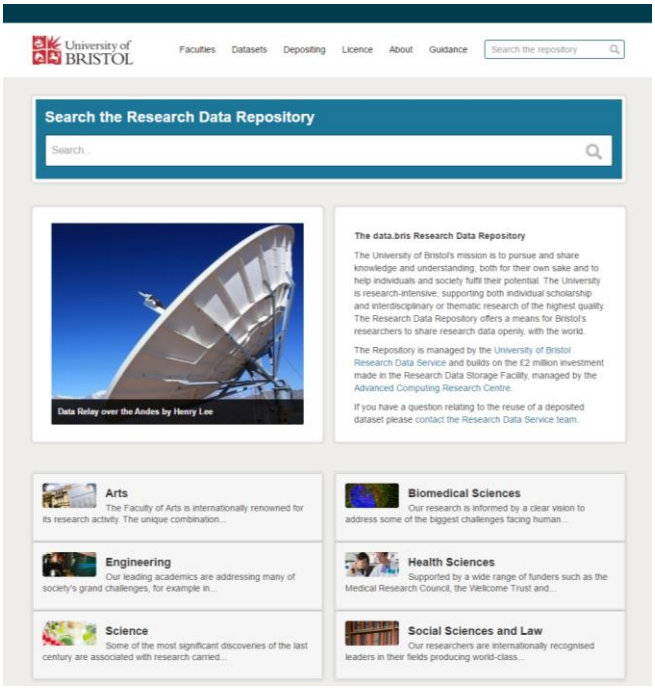
Portal de Dados Abertos Governamentais (Contexto C1)

Data.gov.uk	
http://data.gov.uk	
O portal de dados do Reino Unido foi um dos primeiros portais de dados abertos governamentais a ser lançado (2009). Utiliza o sistema de gestão de conteúdos Drupal para disponibilização do site e o CKAN como sistema de back-end, agregando e disponibilizando dados governamentais de várias fontes, nomeadamente de departamentos centrais do governo britânico, autoridades locais e outros sectores do serviço público.	

Data.gov.au	
http://data.gov.au	
O portal de dados abertos de governação da Austrália, permite a pesquisa, acesso e reutilização de conjuntos de dados públicos do governo Australiano. A equipa deste projeto trabalha transversalmente, em vários governos, de modo a publicar dados e continuar a melhorar a plataforma, baseando-se no feedback dos utilizadores. Para além dos conjuntos de dados, o portal data.gov.au inclui informação acerca de dados por publicar e acerca de dados disponíveis mediante compra.	

Repositório de Dados Científicos (Contexto C2/3)

DART data	
http://dartportal.leeds.ac.uk	
O projeto DART (Detection of Archeological Residues using Remote-sensing Techniques) utiliza o CKAN como repositório de dados. Foi iniciado em 2010 de modo investigar a possibilidade de deteção características arqueológicas em condições tecnicamente exigentes, tendo para tal sido desenvolvida uma ferramenta para a ingestão automática de metadados, enriquecendo significativamente os conjuntos de dados.	

data.bris Research Data Repository	
https://data.bris.ac.uk/data/	
Este repositório é gerido pela Universidade de Bristol, cuja missão é a partilha de conhecimento de modo a usar todo o potencial dos dados gerados, tanto para a instituição em si, como para investigadores externos ou para a sociedade. O repositório teve origem num investimento no Serviço de Investigação da Universidade, correspondente a um investimento inicial de 2 milhões de libras, tendo o mesmo sido gerido pelo Advanced Computing Research Centre. O blog do projeto[22] descreve e discute em detalhe as vantagens do uso do CKAN.	

[22] https://data.blogs.ilrt.org/2012/12/18/ckan-and-data-bris/

CONCLUSÕES

O sistema CKAN constitui um sistema promissor enquanto agregador de dados de investigação e como fonte adicional do portal RCAAP. Sendo um sistema relativamente recente e com uma comunidade muito específica, a sua capacidade de se adaptar a diversos cenários é ainda limitada. Contudo, a vantagem imediata que se vislumbra no sistema é que existem conceitos básicos que são requisitos transversais a portais, repositórios temáticos ou institucionais. O conceito de federação, a apresentação e manuseamento dos dados online assim como a possibilidade de integrar dados de APIs externas são funcionalidades distintivas deste sistema e que se adaptam às novas realidades de interoperabilidade dos sistemas. A usabilidade do sistema é outra mais-valia que deve ser valorizada face a outras opções atualmente disponíveis.

Considerando a utilização deste software em projetos nacionais e comunidades de investigação sedimentadas é previsível a sua continuidade em termos de desenvolvimento e novas funcionalidades, mas sendo um projeto open source, está altamente dependente dos contributos da própria comunidade, da qual o projeto RCAAP poderá também fazer parte.

Existem outras opções dedicadas à gestão e preservação de dados. Exemplo disso são os sistemas DATAVERSE,[23] DSpace ou INVENIO[24] sistemas estes que gerem dados científicos, com estratégias de preservação de dados, e disponibilizando os seus conteúdos online. Além disso, sistemas como o FEDORA[25] poderiam também ser utilizados neste contexto, sendo que neste caso o esforço de desenvolvimento seria muito maior.

Do ponto de vista dos cenários propostos no contexto do projeto RCAAP, o segundo contexto é o que apresenta maior flexibilidade e abrangência para a comunidade, contudo também é o que mais esforço implica. Como alternativa, o primeiro cenário de Portal de dados centralizado, permitindo a integração e agregação de outras fontes de informação de dados científicos é também uma solução adequada e viável que conjuga os esforços num único sistema partilhado tal como já é desenvolvido por exemplo com o Repositório Comum. Além disso, permite autonomia no desenvolvimento das iniciativas locais, permitindo a integração centralizada no serviço do projeto RCAAP. Este contexto permitiria por um lado concretizar um serviço nacional integrado com os sistemas existentes e servir de exemplo para as boas

[23] http://dataverse.org/
[24] http://invenio-software.org/
[25] http://fedorarepository.org/

práticas de gestão de dados, levando a uma maior sensibilização para a gestão de dados científicos no contexto das instituições de ensino e investigação em Portugal.

O terceiro cenário identificado tem como desvantagem o esforço envolvido assim como a ausência de um nível intermédio entre os repositórios e o Portal RCAAP, permitindo a monitorização e controlo.

Contudo, a implementação destes cenários está dependente de decisões políticas, nomeadamente da política nacional de dados científicos, de financiamento disponível e de posicionamento estratégico do projeto RCAAP.